

1/5/1 (Item 1 from file: 351)
DIALOG(R) File 351: Derwent WPI
(c) 2004 Thomson Derwent. All rts. reserv.

012677564 **Image available**
WPI Acc No: 1999-483671/ 199941
XRPX Acc No: N99-360701

Data forwarding procedure in multiprocessor system - involves resending
page where failure occurs during address conversion unit, by referring
two failure page tables

Patent Assignee: NEC CORP (NIDE)
Number of Countries: 001 Number of Patents: 002
Patent Family:

Patent No	Kind	Date	Applicat No	Kind	Date	Week
JP 11203260	A	19990730	JP 986578	A	19980116	199941 B
JP 3237599	B2	20011210	JP 986578	A	19980116	200203

Priority Applications (No Type Date): JP 986578 A 19980116

Patent Details:

Patent No	Kind	Lan	Pg	Main IPC	Filing Notes
JP 11203260	A		16	G06F-015/163	
JP 3237599	B2		16	G06F-015/177	Previous Publ. patent JP 11203260

Abstract (Basic): JP 11203260 A

NOVELTY - An address converter performs conversion between physical
address and logic address. A transmitter (171) transmits data on logic
address to other nodes via communication channel (27). A memory records
page where failure occurs, during address conversion. A resending unit
resends stored page after confirming existence of address conversion
failure with reference to failure page tables (341, 641).

USE - In multiprocessor system.

ADVANTAGE - Data forwarding between nodes is done at high speed.
DESCRIPTION OF DRAWING(S) - The figure shows functional block diagram
of function component of multiprocessor system. (27) Communication
channel; (171) Transmitter; (341, 641) Failure page tables.

Dwg. 1/13

Title Terms: DATA; FORWARDING; PROCEDURE; MULTIPROCESSOR; SYSTEM; PAGE;
FAIL; OCCUR; ADDRESS; CONVERT; UNIT; REFER; TWO; FAIL; PAGE; TABLE

Derwent Class: T01

International Patent Class (Main): G06F-015/163; G06F-015/177

International Patent Class (Additional): G06F-012/10

File Segment: EPI

1/5/2 (Item 1 from file: 347)
DIALOG(R) File 347: JAPIO
(c) 2004 JPO & JAPIO. All rts. reserv.

06261680 **Image available**
MULTIPROCESSOR SYSTEM AND DATA TRANSFER METHOD IN MULTIPROCESSOR SYSTEM

PUB. NO.: 11-203260 A
PUBLISHED: July 30, 1999 (19990730)
INVENTOR(s): SUGAWARA TOMOYOSHI
APPLICANT(s): NEC CORP
APPL. NO.: 10-006578 [JP 986578]
FILED: January 16, 1998 (19980116)
INTL CLASS: G06F-015/163

ABSTRACT

PROBLEM TO BE SOLVED: To perform data transfer between plural nodes at high
speed.

SOLUTION: When a page-out occurs to data to be transferred and a

transmission part 171 or a reception part 271 fails address conversion during data transmission by a communication hardware 27, the transmission part 171 or the reception part 271 demands interruption. An interruption processing 362 records a page which fails in the address translation in a transmission failure page table 341 or a reception failure page table 641, and skips the page to continue data transfer. After a series of data transfer is completed, a process 11 having requested data transmission refers to both of the failure page tables and examines whether or not there is a failure in the address translation. When there is a failure in the address conversion, the process 11 loads the page in a memory 13 and retransmits it.

COPYRIGHT: (C)1999,JPO

(19) 日本国特許庁 (J P)

(12) 特 許 公 報 (B 2)

(11) 特許番号

特許第3237599号
(P3237599)

(45) 発行日 平成13年12月10日 (2001. 12. 10)

(24) 登録日 平成13年10月 5 日 (2001. 10. 5)

(51) Int.Cl.⁷

G 0 6 F 15/177
12/10

識別記号

6 7 6
5 0 7

F I

G 0 6 F 15/177
12/10

6 7 6 A
5 0 7 B

請求項の数 6 (全 16 頁)

(21) 出願番号 特願平10-6578

(22) 出願日 平成10年 1 月16日 (1998. 1. 16)

(65) 公開番号 特開平11-203260

(43) 公開日 平成11年 7 月30日 (1999. 7. 30)

審査請求日 平成10年 1 月16日 (1998. 1. 16)

前置審査

(73) 特許権者 000004237

日本電気株式会社
東京都港区芝五丁目 7 番 1 号

(72) 発明者 菅原 智義
東京都港区芝五丁目 7 番 1 号 日本電気
株式会社内

(74) 代理人 100082935
弁理士 京本 直樹 (外 2 名)

審査官 久保 正典

最終頁に続く

(54) 【発明の名称】 マルチプロセッサシステム及びマルチプロセッサシステムにおけるデータ転送方法

1

(57) 【特許請求の範囲】

【請求項 1】 互いに通信路を介して接続された複数のノードから構成されるマルチプロセッサシステムであって、

前記複数のノードのうちの少なくとも 1 つは、
物理アドレス空間を提供する第 1 の記憶手段と、
論理アドレス空間を提供する第 2 の記憶手段と、
所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行う第 1 のアドレス変換手段と、

前記論理アドレス空間に存在するデータを前記通信路を介して他のノードに送信する第 1 の送信手段と、
前記第 1 の送信手段によって送信される前記データのうち前記第 1 の記憶手段が提供する前記物理アドレス空間に存在しなかったアドレス変換単位に関する情報を記録

2

する第 1 の変換単位情報記録手段と、
前記第 1 の送信手段による前記データの送信が終了した後、前記第 1 の変換単位情報記録手段に記録されている情報に対応するアドレス変換単位のデータを前記通信路を介して前記他のノードに再送する再送手段と、を備え、

前記複数のノードのうちの他の少なくとも 1 つは、
物理アドレス空間を提供する第 3 の記憶手段と、
論理アドレス空間を提供する第 4 の記憶手段と、
所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行う第 2 のアドレス変換手段と、

10

前記複数のノードのうちの少なくとも 1 つから送信されたデータを受信する受信手段と、
前記受信手段が受信したデータを記憶すべき論理アドレ

3

スに対応する物理アドレスが前記第3の記憶手段に存在しないアドレス変換単位に関する情報を記録する第2の変換単位情報記録手段と、を備え、

前記複数のノードのうちの少なくとも1つは、

前記通信路を介して前記第2の変換単位情報記録手段に記録された情報を読み出す変換単位情報読み出し手段と、

前記変換単位情報読み出し手段が読み出した情報に対応するデータを前記第2のアドレス変換手段にアドレス変換させて前記第4の記憶手段から前記第3の記憶手段に書き込む手段とを備え、

前記再送手段は、前記変換単位情報読み出し手段が読み出した情報に対応するアドレス変換単位のデータをさらに前記複数のノードのうちの少なくとも1つから前記通信路を介して前記複数のノードのうちの他の少なくとも1つに再送することを特徴とするマルチプロセッサシステム。

【請求項2】前記受信手段は、受信したデータのうち前記第3の記憶手段が提供する前記物理アドレス空間に存在しなかったアドレス変換単位のデータを破棄することを特徴とする請求項1に記載のマルチプロセッサシステム。

【請求項3】前記送信されるデータのうちの最後の送信単位には、該送信単位が最後のものであることを示す情報が含まれ、

前記複数のノードのうちの他の少なくとも1つは、

前記受信手段が前記送信単位のうちの最後のデータを受信したときに、前記前記第2の変換単位情報記録手段に記録されている情報を参照する第2の変換単位情報参照手段と、

前記第2の変換単位情報参照手段による参照の結果、いずれかのアドレス変換単位を記憶すべき論理アドレスに対応する物理アドレスが前記第3の記憶手段に存在しなかったことを示すときに、前記最後の送信単位を含むアドレス変換単位に関する情報を前記第2の変換単位情報記録手段に記録させる第2の変換単位情報制御手段と、前記第2の変換単位情報参照手段による参照の結果、いずれかのアドレス変換単位を記憶すべき論理アドレスに対応する物理アドレスが前記第3の記憶手段に存在しなかったことを示すときに、前記最後の送信単位のデータを破棄する破棄手段とをさらに備えることを特徴とする請求項1または2に記載のマルチプロセッサシステム。

【請求項4】前記複数のノードのうちのさらに他の少なくとも1つは、

前記第1の記憶手段が提供する物理アドレス空間に対する論理アドレス空間をさらに提供する第5の記憶手段と、

前記複数のノードのうちの少なくとも1つからの要求に従うアドレス変換単位を送信する第2の送信手段とを備え、

4

前記複数のノードのうちの少なくとも1つは、

前記第1の送信手段による前記データの送信が終了した後、前記第1の変換単位情報記録手段に記録されている情報に対応するアドレス変換単位が前記第5の記憶手段に記憶されているかどうかを判定する判定手段と、

前記判定手段による判定の結果、前記第5の記憶手段に記憶されていると判定されたアドレス変換単位を前記第2の送信手段に送信させるべき前記要求を前記通信路を介して行う送信要求手段と、をさらに備えることを特徴とする請求項1乃至3のいずれか1項に記載のマルチプロセッサシステム。

【請求項5】前記再送手段は、

前記第1の変換単位情報記録手段に記録されている情報に対応するデータを前記第1のアドレス変換手段にアドレス変換させて前記第2の記憶手段から前記第1の記憶手段に書き込む手段を備え、

該手段によって前記第1の記憶手段に書き込まれたデータを前記通信路を介して前記他のノードに再送することを特徴とする請求項1乃至4のいずれか1項に記載のマルチプロセッサシステム。

【請求項6】互いに通信路を介して接続され、それぞれ物理アドレス空間を提供する第1の記憶装置と、論理アドレス空間を提供する第2の記憶装置と、所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行うアドレス変換機構とを有する複数のノードから構成されるマルチプロセッサシステムにおけるデータ転送方法であって、

前記複数のノードのうちの少なくとも1つから、該ノードの論理アドレス空間に存在するデータを前記通信路を介して他のノードに送信する送信ステップと、

前記送信ステップで送信される前記データのうち前記第1の記憶装置が提供する前記物理アドレス空間に存在しなかったアドレス変換単位に関する情報を記録する第1の変換単位情報記録ステップと、

前記送信ステップでの前記データの送信が終了した後、前記第1の変換単位情報記録ステップで記録した情報に対応するアドレス変換単位のデータを前記複数のノードのうちの少なくとも1つから前記通信路を介して前記他のノードに再送する再送ステップと、

前記送信ステップで前記複数のノードの少なくとも1つから送信されたデータを受信した前記他のノードが、該データを記憶すべき論理アドレスに対応する物理アドレスが前記第1の記憶装置に存在しないアドレス変換単位に関する情報を記録する第2の変換単位情報記録ステップとを含み、

前記再送ステップは、前記第2の変換単位情報記録ステップで記録した情報に対応するアドレス変換単位のデータをさらに前記複数のノードのうちの少なくとも1つから前記通信路を介して前記他のノードに再送することを特徴とするマルチプロセッサシステムにおけるデータ転

送方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、マルチプロセッサシステム及びマルチプロセッサシステムにおけるデータ転送方法に関し、特に各ノードの通信ハードウェアが仮想アドレスによるデータ転送機能をサポートするマルチプロセッサシステムに適用するものに関する。

【0002】

【従来の技術】従来、ネットワークを介して複数のノードが結合されたマルチプロセッサシステムにおいて、ノード間でデータ転送を行う場合には、オペレーティングシステム（OS）の機能を利用する方法が一般的であった。しかしながら、この方法では、OSをシステムコールするための通信遅延時間が大きくなり、また、ユーザ空間とシステム空間との間でデータをコピーしなければならないため、データの転送速度が遅くなるという問題があった。

【0003】そこで、近年の分散メモリ型並列マシンやクラスタ型マルチプロセッサシステムでは、ノード間通信の機能として仮想アドレスによるデータ転送をサポートするものも用いられている。このようなシステムでは、ユーザプロセスはOSを介することなく通信ハードウェアによるデータ転送を可能としている。しかしながら、このようなシステムでは、送信されるデータや受信したデータを格納すべきバッファは、主記憶装置上に存在する保証がなく、アドレス変換に失敗して通信ハードウェアの動作が停止する可能性があった。

【0004】この問題を解決するため、特開平6-19856号公報において、予め通信領域を特定し、その部分はページアウトされないようにOSに登録しておくことによって、データの転送時にアドレス変換の失敗が生じないようにする方法（第1の方法）が提案されている。

【0005】また、アドレス変換の失敗があった時点でデータの送信あるいは受信を停止して、通信ハードウェアから割り込みをかけてオペレーティングシステムを呼び出し、データの送信或いは受信を一旦停止して、ページフォルトに対する処理を行ってからデータの送信或いは受信を再開する方法（第2の方法）も提案されている。

【0006】

【発明が解決しようとする課題】しかしながら、上記第1の方法では、ユーザプロセス中の任意のデータを転送するためには、結局データを通信領域にコピーする必要があり、データ転送の速度を低下させるという問題があった。また、上記第2の方法では、ページフォルトの処理の間、データの送信或いは受信が待たされることとなるので、すべてのデータについての転送を終了するまでに時間がかかるという問題があった。

【0007】本発明は、上記従来例の問題点を解消するためになされたものであり、複数のノード間でのデータ転送を高速に行うことができるマルチプロセッサシステム及びこのようなマルチプロセッサシステムにおけるデータ転送方法を提供することを目的とする。

【0008】

【課題を解決するための手段】上記目的を達成するため、本発明の第1の観点にかかるマルチプロセッサシステムは、互いに通信路を介して接続された複数のノードから構成されるマルチプロセッサシステムであって、前記複数のノードのうちの少なくとも1つは、物理アドレス空間を提供する第1の記憶手段と、論理アドレス空間を提供する第2の記憶手段と、所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行う第1のアドレス変換手段と、前記論理アドレス空間に存在するデータを前記通信路を介して他のノードに送信する第1の送信手段と、前記第1の送信手段によって送信される前記データのうち前記第1の記憶手段が提供する前記物理アドレス空間に存在しなかったアドレス変換単位に関する情報を記録する第1の変換単位情報記録手段と、前記第1の送信手段による前記データの送信が終了した後、前記第1の変換単位情報記録手段に記録されている情報に対応するアドレス変換単位のデータを前記通信路を介して前記他のノードに再送する再送手段と、を備え、前記送信されるデータのうちの最後の送信単位には、該送信単位が最後のものであることを示す情報が含まれ、前記複数のノードのうちの少なくとも1つは、前記第1の送信手段が前記送信単位のうちの最後のデータを送信するときに、前記第1の変換単位情報記録手段に記録されている情報を参照する第1の変換単位情報参照手段と、前記第1の変換単位情報参照手段による参照の結果、いずれかのアドレス変換単位が前記第1の記憶手段が提供する前記物理アドレス空間に存在しなかったことを示すときに、前記最後の送信単位を含むアドレス変換単位に関する情報を前記第1の変換単位情報記録手段に記録させる第1の変換単位情報制御手段と、前記第1の変換単位情報参照手段による参照の結果、いずれかのアドレス変換単位が前記第1の記憶手段が提供する前記物理アドレス空間に存在しなかったことを示すときに、前記第1の送信手段による最後の送信単位のデータの送信を停止させる送信停止手段とをさらに備えることを特徴とする。

【0009】

【0010】

【0011】上記マルチプロセッサシステムでは、前記送信手段によって送信されるデータのうちの物理アドレス空間に存在しないアドレス変換単位、例えば、前記第2の記憶手段にページアウトされているページは、それに関する情報（例えば、ページ番号）が前記第1の変換単位情報記録手段に記録された後、前記再送手段によって再

送される。このため、前記変換情報記録手段への記録の際にページイン要求を出し、それと並行してデータ処理を続行させること（プリページング）が可能となる。これにより、ページイン処理とデータ転送とをオーバーラップさせることができ、全体としてデータ転送を高速化することができる。

【0012】また、前記再送手段は、前記第1の変換単位情報記録手段に記録されている情報に従ってアドレス変換単位を再送すればよい。また、前記第1のアドレス変換手段は、前記第1の変換単位情報記録手段に記録されている情報に従ってアドレス変換単位のアドレス変換を行えばよい。このため、必要のないアドレス変換やデータの再送をしなくてもよく、無駄を防ぐことができる。そして、最後の送信単位は、前記再送手段による再送があっても、全体のデータの中で必ず最後に送信されることとなるので、例えば、データ中の最後の1ビットを受信完了を示すフラグとして用いるプログラムでも、正しく動作させることが可能となる。

【0013】

【0014】

【0015】上記目的を達成するため、本発明の第2の観点にかかるマルチプロセッサシステムは、互いに通信路を介して接続された複数のノードから構成されるマルチプロセッサシステムであって、前記複数のノードのうちの少なくとも1つは、物理アドレス空間を提供する第1の記憶手段と、論理アドレス空間を提供する第2の記憶手段と、所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行う第1のアドレス変換手段と、前記論理アドレス空間に存在するデータを前記通信路を介して他のノードに送信する第1の送信手段と、前記第1の送信手段によって送信される前記データのうち前記第1の記憶手段が提供する前記物理アドレス空間に存在しなかったアドレス変換単位に関する情報を記録する第1の変換単位情報記録手段と、前記第1の送信手段による前記データの送信が終了した後、前記第1の変換単位情報記録手段に記録されている情報に対応するアドレス変換単位のデータを前記通信路を介して前記他のノードに再送する再送手段と、を備え、前記複数のノードのうちの他の少なくとも1つは、物理アドレス空間を提供する第3の記憶手段と、論理アドレス空間を提供する第4の記憶手段と、所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行う第2のアドレス変換手段と、前記複数のノードのうちの少なくとも1つから送信されたデータを受信する受信手段と、前記受信手段が受信したデータを記憶すべき論理アドレスに対応する物理アドレスが前記第3の記憶手段に存在しないアドレス変換単位に関する情報を記録する第2の変換単位情報記録手段と、を備え、前記複数のノードのうちの少なくとも1つは、前記通信路を介して前記第2の変換単位情報記録手段に記録された情報

を読み出す変換単位情報読み出し手段と、前記変換単位情報読み出し手段が読み出した情報に対応するデータを前記第2のアドレス変換手段にアドレス変換させて前記第4の記憶手段から前記第3の記憶手段に書き込む手段とを備え、前記再送手段は、前記変換単位情報読み出し手段が読み出した情報に対応するアドレス変換単位のデータをさらに前記複数のノードのうちの少なくとも1つから前記通信路を介して前記複数のノードのうちの他の少なくとも1つに再送することを特徴とする。

10 【0016】このように構成した場合、上記マルチプロセッサシステムにおいて、前記受信手段は、受信したデータのうち前記第3の記憶手段が提供する前記物理アドレス空間に存在しなかったアドレス変換単位のデータを破棄するものとしてすることができる。

【0017】この場合、受信側となるノードで受信したアドレス変換単位を記憶すべき物理アドレスが前記第3の記憶手段にないとき、例えば、受信したデータを記憶すべきページのうち前記第4の記憶手段にページアウトされているページは、それに関する情報（例えば、ページ番号）が前記第2の変換情報記録手段に記録された後、前記再送手段によって再送される。このため、前記変換情報記録手段への記録の際にページイン要求を出し、それと並行してデータ処理を続行させること（プリページング）が可能となる。これにより、ページイン処理とデータ転送とをオーバーラップさせることができ、全体としてデータ転送を高速化することができる。

【0018】このように構成した場合にも、また、前記送信されるデータのうちの最後の送信単位には、該送信単位が最後のものであることを示す情報が含まれ、前記複数のノードのうちの他の少なくとも1つは、前記受信手段が前記送信単位のうちの最後のデータを受信したときに、前記前記第2の変換単位情報記録手段に記録されている情報を参照する第2の変換単位情報参照手段と、前記第2の変換単位情報参照手段による参照の結果、いずれかのアドレス変換単位を記憶すべき論理アドレスに対応する物理アドレスが前記第3の記憶手段に存在しなかったことを示すときに、前記最後の送信単位を含むアドレス変換単位に関する情報を前記第2の変換単位情報記録手段に記録させる第2の変換単位情報制御手段と、前記第2の変換単位情報参照手段による参照の結果、いずれかのアドレス変換単位を記憶すべき論理アドレスに対応する物理アドレスが前記第3の記憶手段に存在しなかったことを示すときに、前記最後の送信単位のデータを破棄する破棄手段とをさらに備えるものとしてすることができる。

【0019】この場合、最後の送信単位は、前記再送手段による再送があっても、全体のデータの中で必ず最後に送信されることとなるので、例えば、データ中の最後の1ビットを受信完了を示すフラグとして用いるプログラムでも、正しく動作させることが可能となる。

【0020】上記第1、第2の観点にかかるマルチプロセッサシステムにおいて、前記複数のノードのうちのさらに他の少なくとも1つは、前記第1の記憶手段が提供する物理アドレス空間に対する論理アドレス空間をさらに提供する第5の記憶手段と、前記複数のノードのうちの少なくとも1つからの要求に従うアドレス変換単位を送信する第2の送信手段とを備え、前記複数のノードのうちの少なくとも1つは、前記第1の送信手段による前記データの送信が終了した後、前記第1の変換単位情報記録手段に記録されている情報に対応するアドレス変換単位が前記第5の記憶手段に記憶されているかどうかを判定する判定手段と、前記判定手段による判定の結果、前記第5の記憶手段に記憶されていると判定されたアドレス変換単位を前記第2の送信手段に送信させるべき前記要求を前記通信路を介して行う送信要求手段と、をさらに備えるものとすることができる。

【0021】前記第5の記憶手段に送信すべきデータのアドレス変換単位が記憶されているときは、その第5の記憶手段を有する前記複数のノードのうちのさらに他の少なくとも1つから前記複数のノードのうちの少なくとも1つにそのアドレス変換単位が送信されることとなるので、データの再送のために必要となる時間を短縮することができる。さらに、送信側となる前記複数のノードのうちの少なくとも1つにおける負担を軽減することができる。

【0022】上記第1、第2の観点にかかるマルチプロセッサシステムにおいて、前記再送手段は、前記第1の変換単位情報記録手段に記録されている情報に対応するデータを前記第1のアドレス変換手段にアドレス変換させて前記第2の記憶手段から前記第1の記憶手段に書き込む手段を備え、該手段によって前記第1の記憶手段に書き込まれたデータを前記通信路を介して前記他のノードに再送するものとすることができる。

【0023】上記目的を達成するため、本発明の第3の観点にかかるマルチプロセッサシステムにおけるデータ転送方法は、互いに通信路を介して接続され、それぞれ物理アドレス空間を提供する第1の記憶装置と、論理アドレス空間を提供する第2の記憶装置と、所定のアドレス変換単位毎に物理アドレスと論理アドレスとの間のアドレス変換を行うアドレス変換機構とを有する複数のノードから構成されるマルチプロセッサシステムにおけるデータ転送方法であって、前記複数のノードのうちの少なくとも1つから、該ノードの論理アドレス空間に存在するデータを前記通信路を介して他のノードに送信する送信ステップと、前記送信ステップで送信される前記データのうち前記第1の記憶装置が提供する前記物理アドレス空間に存在しなかったアドレス変換単位に関する情報を記録する第1の変換単位情報記録ステップと、前記送信ステップでの前記データの送信が終了した後、前記第1の変換単位情報記録ステップで記録した情報に対応

するアドレス変換単位のデータを前記複数のノードのうちの少なくとも1つから前記通信路を介して前記他のノードに再送する再送ステップと、前記送信ステップで前記複数のノードの少なくとも1つから送信されたデータを受信した前記他のノードが、該データを記憶すべき論理アドレスに対応する物理アドレスが前記第1の記憶装置に存在しないアドレス変換単位に関する情報を記録する第2の変換単位情報記録ステップとを含み、前記再送ステップは、前記第2の変換単位情報記録ステップで記録した情報に対応するアドレス変換単位のデータをさらに前記複数のノードのうちの少なくとも1つから前記通信路を介して前記他のノードに再送することを特徴とする。

【0024】

【発明の実施の形態】以下、添付図面を参照して、本発明の実施の形態について説明する。

【0025】【第1の実施の形態】図1は、この実施の形態にかかるマルチプロセッサシステムの機能構成を示す機能ブロック図である。図示するように、このマルチプロセッサシステムは、ネットワーク10を介して互いに結合された複数のノード3-1～3-nからなる。

【0026】ノード3-1～3-nは、同様の機能構成を有し、それぞれCPU（図示せず）、メモリ13、二次記憶装置18及び通信ハードウェア27を備える。また、ノード3-1～3-nのそれぞれの上では、プロセス11及びカーネル36が動作する。このマルチプロセッサシステムで結合することが可能なノード3-1～3-nは、最大で256個とすることができる。

【0027】メモリ13は、ノード3-1～3-nのそれぞれの主記憶装置であり、物理アドレス空間を提供する。二次記憶装置18は、磁気ディスク装置などによって構成され、ノード3-1～3-nのそれぞれに仮想アドレス空間を提供する。

【0028】通信ハードウェア27は、例えば、データの最大転送サイズを256KBとし、ページサイズを4KBとするページング方式で転送されるデータの仮想アドレスを物理アドレスに変換する機能を有する。通信ハードウェア27は、また、割り込み発生部172、送信部171、受信部272及びリモートリード部571の各機能を有する。通信ハードウェア27は、さらにプロセス11からのデータの送受信の要求を書き込むための要求領域176と、要求されたデータの送受信が終了したことを示す情報を書き込むための確認領域175とを有する。要求領域176及び確認領域175は、仮想空間内にマップされている。

【0029】割り込み発生部172は、アドレス変換に失敗したとき、すなわち送信或いは受信すべきデータについてページフォルトが発生したときに、CPUに割り込みを要求し、カーネル36に割り込み処理をさせる。送信部171は、メモリ13に書き込まれているデータ

を読み出し、パケットに分割してネットワーク10に送出する。受信部271は、ネットワーク10から受け取ったパケット中のデータをメモリ13に書き込む。リモートリード部571は、ネットワーク10から受け取ったリモートリード要求に応答して、カーネル36の受信失敗ページ表641などを他のノードに読み込ませる機能を提供する。

【0030】カーネル36は、カーネル間通信や仮想記憶管理などの基本機能を提供するほか、割り込み処理362及び遠隔ページアクセス761の機能を提供する。カーネル36は、また、仮想アドレス空間と物理アドレス空間との間でのアドレス変換を行うためのページテーブル40と、データの送信時及び受信時にそれぞれページフォルトが発生したページに関する情報を記録するための送信失敗ページ表341と受信失敗ページ表641とを有する。

【0031】割り込み処理362の機能は、CPUに割り込みが生じたときにその割り込み要因を特定し、特定した割り込み要因が送信部171或いは受信部271におけるアドレス変換の失敗（ページフォルト）である場合には、アドレス変換に失敗したページに関する情報を送信失敗ページ表341及び受信失敗ページ表641にそれぞれ記憶する。割り込み処理362では、送信失敗ページ表341或いは受信失敗ページ表641にアドレス変換に失敗したページに関する情報を記録した後、要求領域176にデータの送受信の再開要求を書き込むことにより、送信部171或いは受信部271の動作を再開させる。

【0032】送信失敗ページ表341は、送信部171が送信するデータについてページフォルトが発生したときに、そのアドレス変換に失敗したページに関する情報を記録する。受信失敗ページ表641は、受信部271が受信するデータについてページフォルトが発生したときに、そのアドレス変換に失敗したページに関する情報を記録する。送信失敗ページ表341と受信失敗ページ表641とは、通信ハードウェア27で転送可能なデータの最大サイズが決まっているため、図2に示すように、仮想アドレス+ビット列でアドレス変換に失敗したページを記録する。

【0033】これらの失敗ページ表では、図2に示すように、1ワード目には最初にアドレス変換に失敗したページのページ番号PNUMが入れられる。2ワード目以降には、各ビットの値で対応するページがアドレス変換に失敗したかどうかを示し、例えば、2ワード目の第31ビットの値が「1」であるならば、PNUM+31ページでアドレス変換に失敗したことを、3ワード目の第31ビットの値が「1」であるならば、PNUM+63ページでアドレス変換に失敗したことを示すものである。なお、PNUMの初期値は、「-1」とする。

【0034】この実施の形態のマルチプロセッサシステ

ムでは、データの送信側となるプロセス11は、他の複数のノードのプロセスに対して同時にデータ転送を行うことはない。このため、ノード3-1~3-nのそれぞれにおいて、送信失敗ページ表341は1つだけ用意されている。一方、プロセス11がデータの受信側となるときには、複数の他のノードのプロセスからデータを同時に受信することがあるので、受信失敗ページ表641は、送信側となるプロセスの数だけ用意される。なお、これらの送信失敗ページ表341及び受信失敗ページ表641の仮想空間上の配置を、図3に示す。

【0035】遠隔ページアクセス761は、当該ノードとネットワーク10を介して接続された他のノードにあるプロセスの仮想空間にアクセスし、アクセスしたページがページアウトされているときには、そのノードのカーネルにアクセスしたページをページインさせる。

【0036】プロセス11は、通信ハードウェア27の要求領域176に送信要求を書き込むことによって、データ12を他のノードのプロセスに送信する。プロセス11は、通信ハードウェア27の確認領域175の値を調べることによって、データ12の送信が終了したことを確認する。プロセス11は、他のノードから受信したデータを書き込むための受信バッファ22を用意している。

【0037】プロセス11は、また、送信失敗ページ表341や受信失敗ページ表641を参照して、データの送受信時にアドレス変換の失敗があったかどうかを確認し、アドレス変換の失敗があったページについて再送処理を行う。なお、プロセス11には、システム中でプロセスを一意に定められるように、「0」から「255」までのいずれかの識別子が割り付けられる。なお、データ12及び受信バッファ22は、仮想空間にあるもので、メモリ13上に存在せずに、二次記憶装置18にページアウトされていることがあり得る。

【0038】以下、この実施の形態のマルチプロセッサシステムにおける動作について説明する。以下の説明において、ノード3-1をデータの送信側とし、ノード3-nをデータの受信側とする。ここで、ノード3-1のプロセス11を送信側プロセス11と呼び、ノード3-nのプロセス11を受信側プロセス11と呼ぶこととする。

【0039】まず、ノード3-1とノード3-nとの間において、データ転送が開始される前に、ノード3-1とノード3-nとのそれぞれにおいて実行される準備処理について説明する。

【0040】カーネル36は、プロセス11を生成したときに、生成したプロセス11に対応するページテーブル40と送信失敗ページ表341と受信失敗ページ表641を記憶する領域をメモリ13上に確保し、これらをページアウトしないよう指定する。カーネル36は、さらに送信失敗ページ表341と受信失敗ページ表641

とを、それぞれ仮想アドレスSND_FTBL_ADDRとRCV_FTBL_ADDRにマップする。また、プロセス11は、カーネル36のマップ機能を用いて、要求領域176と確認領域175とを仮想空間内にマップする。

【0041】次に、送信側ノード3-1から受信側ノード3-nへのデータ転送の大まかな流れについて、図4のフローチャートを参照して説明する。送信側プロセス11は、まず、自プロセスの送信失敗ページ表341を初期化する。送信側プロセス11は、さらに、通信ハードウェア27の送信機能を利用して、受信側プロセス11の受信失敗ページ表のうち送信側プロセス11のものに対応する部分を初期化する。すなわち、送信側プロセス11は、要求領域176に所定の送信要求を書き込んで、受信失敗ページ表641の初期値を受信側ノード3-nの該当領域に書き込む（ステップS11）。なお、該当領域の計算は、数式1に従って行われる。

【数1】RCV_FTBL_ADDR+3ワード×プロセスの識別子

【0042】送信側プロセス11は、通信ハードウェア27の要求領域176に所定の送信要求を書き込むことにより、通信ハードウェア27にデータ12の送信を要求する（ステップS12）。

【0043】送信側ノード3-1の送信部171は、要求領域176に書き込まれた送信要求を読み出すことによって、ネットワーク10を介してのノード3-nへのデータ12の送信を開始する。データ12は、前述したようにパケットの形式で行われる。送信部171は、データ12の途中でアドレス変換に失敗した場合には、データ12の送信を停止して、CPUに対して割り込みを要求する（ステップS13）。

【0044】ステップS13で割り込みが発生すると、送信側ノード3-1のカーネル36は、割り込み処理を行い、割り込み処理の終了の後、送信部171にデータ12の送信を続行させる（ステップS14）。なお、この割り込み処理については、さらに詳しく後述する。

【0045】受信側ノード3-nの受信部271は、ネットワーク10から受け取ったデータを受信バッファ22に書き込む。受信部271は、データの書き込みの途中でアドレス変換に失敗した場合には、データの受信を停止して、CPUに対して割り込みを要求する（ステップS15）。

【0046】ステップS15で割り込みが発生すると、受信側ノード3-nのカーネル36は、割り込み処理を行い、割り込み処理の終了の後、送信部271にデータ12の送信を続行させる（ステップS16）。なお、この割り込み処理については、さらに詳しく後述する。

【0047】一方、送信側プロセス11は、送信側ノード3-1の確認領域175を参照することによって、要求したデータ12の送信が終了したことを確認する（ス

テップS17）。

【0048】送信側プロセス11は、データ12の送信が終了したことを確認すると、さらに送信側ノード3-1の送信失敗ページ表341と、受信側ノード3-nの受信失敗ページ表641とを参照して、アドレス変換の失敗があったかどうかをチェックする（ステップS18）。ここで、送信側ノード3-1の送信部171でのアドレス変換の失敗は、通常のメモリアクセスで送信失敗ページ表341を読み、PNUMの値が「-1」以外であれば発生していることがわかる。受信側ノード3-nの受信部271でのアドレス変換の失敗は、リモートリード部571を利用してリモートリードで受信側ノード3-nの受信失敗ページ表641を読み、PNUMの値が「-1」以外であれば発生していることがわかる。

【0049】ステップS18で送信側ノード3-1の送信部171或いは受信側ノード3-nの受信部271のいずれかでアドレス変換の失敗があったことが確認された場合には、送信側プロセス11は、必要に応じてデータ12の再送処理を行う（ステップS19）。なお、この再送処理については、さらに詳しく後述する。

【0050】次に、送信側ノード3-1及び受信側ノード3-nのカーネル36の割り込み処理362に機能が実行する割り込み処理（ステップS14、S16）について、図5のフローチャートを参照して詳しく説明する。割り込みが発生した場合、カーネル36の割り込み処理40の機能は、その割り込み要因について調べる（ステップS21）。

【0051】ステップS21で調べた割り込み要因が送信部171におけるアドレス変換の失敗である場合には、この例では、送信側ノード3-1で生じた割り込みである。この場合は、データ12の送信が停止されると共に、アドレス変換に失敗したページの仮想アドレスに関する情報が送信側ノード3-1の送信失敗ページ表341に記録される（ステップS22）。このとき、送信側ノード3-1の割り込み処理部362は、当該ページのページイン要求を出し、メモリ13にプリページングさせることができる。

【0052】送信失敗ページ表341への記録が終了すると、データ12の送信を再開させるための所定の送信要求が通信ハードウェア27の要求領域176に書き込まれ、送信側ノード3-1の送信部171は、当該アドレス変換に失敗したページの次のページから、データ12の送信を再開する。但し、アドレス変換に失敗したページが送信すべきデータ12の最後のページであった場合には、データ12の送信処理は終了する（ステップS23）。

【0053】ステップS21で調べた割り込み要因が受信部271におけるアドレス変換の失敗である場合には、この例では、受信側ノード3-nで生じた割り込みである。この場合は、データ12の受信が停止されると

共に、アドレス変換に失敗したページの仮想アドレスに関する情報が受信側ノード3-nの受信失敗ページ表641に記録される(ステップS24)。このとき、このとき、受信側ノード3-nの割り込み処理部362は、当該ページのページイン要求を出し、メモリ13にプリページングさせることができる。

【0054】受信失敗ページ表641への記録が終了すると、受信側ノード3-nの受信部271は、当該アドレス変換に失敗したページを破棄し、次のページからのデータ12を受信する。

【0055】なお、ステップS21で調べた割り込み要因がアドレス変換の失敗以外である時は、図5のフローチャートには示さないが、その割り込み要因に従った割り込み処理が行われる。

【0056】次に、送信側プロセス11が実行する再送処理(ステップS19)について、図6(a)、(b)のフローチャートを参照して詳しく説明する。送信側ノード3-1でのアドレス変換の失敗があった場合には、図6(a)に示すように、送信側プロセス11は、送信側ノード3-1の送信側失敗ページ表341を参照して、アドレス変換に失敗したページの仮想アドレスを計算する(ステップS31)。

【0057】次に、送信側プロセス11は、ステップS31で計算した仮想アドレスにアクセスしてページフォルトを起こさせ、カーネル36に当該アドレス変換に失敗したページを二次記憶装置18からメモリ13にページインさせる(ステップS32)。そして、送信側プロセス11は、そのページインされたページを再送する(ステップS33)。

【0058】一方、受信側ノードでのアドレス変換の失敗があった場合には、図6(b)に示すように、送信側プロセス11は、リモートリードで呼んだ受信側ノード3-nの受信失敗ページ表641を参照して、アドレス変換に失敗したページの受信側ノード3-nにおける仮想アドレスを計算する(ステップS41)。

【0059】次に、送信側プロセス11は、カーネル36の遠隔ページアクセス761の機能を利用して、受信側ノード3-nのステップS41で計算した仮想アドレスにアクセスしてページフォルトを起こさせ、受信側ノード3-nのカーネル36に当該アドレス変換に失敗したページを二次記憶装置18からメモリ13にページインさせる(ステップS42)。そして、送信側プロセス11は、当該受信側ノード3-nのメモリ13に書き込めなかったデータを再送する。

【0060】上記の処理で、送信側ノード3-1或いは受信側ノード3-2でアドレス変換に失敗したページが複数ある場合には、送信側プロセス11は、そのすべてのページについてステップS31、S32、S41、S42の処理を行った上で、ステップS33、S43の処理を行う。

【0061】なお、カーネル36は、物理空間の空き容量が不足した場合には、ページアウトの処理を行うが、通信ハードウェア27がページアウトされるページを参照している可能性もある。このため、カーネル36がページアウトの処理を行い、ページテーブル40に変更を加える場合には、通信ハードウェアの動作を一旦停止させて、通信ハードウェア27の処理によってアドレス変換が行われないようにする。そして、ページテーブル40の内容が変更された後に、通信ハードウェア27の動作を再開させる。

【0062】以下、図7を参照して、この実施の形態にかかるマルチプロセッサシステムの具体的な動作例について説明する。この例では、転送されるデータ12は4Kバイトのページを4ページ含み、そのうち、1ページ目と4ページ目がメモリ13上に存在し、2ページ目と3ページ目が二次記憶装置18にページアウトされているものとする。

【0063】プロセス11がデータ12を送信しようとした場合、1ページ目121の先頭アドレスは物理アドレス0x10000に変換される。プロセス11は、その物理アドレスから1ページ分(0x1000バイト)のデータを送信部171に転送させる。このとき、ステップS13のデータ送信では、アドレス変換の失敗はない。

【0064】しかし、2ページ目122(0x504000~0x507FFF)はページアウトされているので、アドレス変換は失敗して割込み発生部172により割り込みが発生する。これにより、ステップS13のデータ送信でアドレス変換の失敗があったことがカーネル36に通知される。

【0065】割り込み処理362ではまず、送信失敗ページ表341にこのページのページ番号0x504を1ワード目に記録する。さらに、送信失敗ページ表341の2ワード目の第0ビットを「1」にする(ステップS22)。このアドレス変換に失敗したページは、とばされて送信が進む(ステップS23)。同様に、3ページ目もページアウトされているので送信失敗ページ表341の2ワード目の第1ビットも「1」とされる。

【0066】送信側プロセス11は、データ12の送信終了後、失敗ページ表341を調べる(ステップS18)。すると、送信側プロセス11は、送信側で第2ページと第3ページにアドレス変換の失敗があったことがわかるので(ステップS31)、これらのページにアクセスしてページフォルトを起こさせ、カーネル36に第1ページと第2ページとを二次記憶装置18からメモリ13にページインをさせる(ステップS32)。そして、送信側プロセス11は、ページインが終了した後で失敗した第2ページと第4ページのデータを再送する(ステップS33)。

【0067】以上説明したように、この実施の形態のマ

ルチプロセッサシステムでは、送信プロセス11が送信失敗ページ表341或いは受信失敗ページ表641を参照してアドレス変換に失敗したページの有無をチェックした上でページインや再送処理を行うことができる。このため、送信プロセス11が必要ないと判断したときは、ページインや再送処理を行わなくて済むので、無駄なページインの処理や再送処理を行わなくても済む。

【0068】また、この実施の形態のマルチプロセッサシステムでは、転送されるデータについてアドレス変換に失敗したときに、ページインの要求だけを出しておき、それと並行してデータ転送を継続することが可能である。これにより、ページイン処理とデータ転送とをオーバーラップさせることができ、全体としてデータ転送を高速化することができる。

【0069】〔第2の実施の形態〕この実施の形態では、データ転送を複数回行なった後で一回だけ失敗確認と再送をできるように失敗ページ表を変更することにより、失敗確認の回数を減らすことができるマルチプロセッサシステムについて説明する。

【0070】この実施の形態のマルチプロセッサシステムの構成は、第1の実施の形態のもの（図1）とほぼ同一である。但し、この実施の形態のマルチプロセッサシステムでは、アドレス変換に失敗したページの仮想アドレスを記録するためのデータ構造として、第1の実施の形態の失敗ページ表の代わりに、失敗ページカウンタと失敗ページ表ポインタを使用する。

【0071】図8は、失敗ページカウンタの構成を示す図である。失敗ページカウンタは、アドレス変換に失敗したページの数を保存するカウンタと、カーネルが失敗ページを記録するかどうかを指定するため書き込み許可フラグとから構成される。カウンタは31ビットで構成され、データの転送を開始する前に送信プロセスによって「0」に初期化され、アドレス変換の失敗が起こる度にカーネル36によって「1」ずつインクリメントされる。また、書き込み許可フラグは1ビットで構成されている。失敗ページを記録していないとき、すなわちデータ12の送受信の終了後に失敗ページの確認の必要がない場合には、フラグは「0」に設定される。失敗ページを記録する場合には、フラグは「1」に設定される。

【0072】失敗ページ表ポインタは、アドレス変換に失敗したページの仮想アドレスを保持するための領域を指し示すポインタである。送信プロセス11あるいは受信プロセス11は、この仮想アドレスの記録領域をデータ転送の前に確保し、失敗ページ表ポインタに、その記録領域の先頭アドレスを設定する。

【0073】このような失敗ページカウンタと失敗ページ表ポインタとからなるデータ構造は、プロセス毎にそれぞれ用意される。但し、第1の実施の形態の場合と異なり、受信用にはプロセス識別子の数だけこのようなデータ構造を用意すればよい。なお、プロセス11の仮想

空間における上記のデータ構造の配置の例を、図9に示す。

【0074】以下、この実施の形態のマルチプロセッサの動作について説明する。カーネル36は、プロセス11を生成する際に、送信用および受信用の失敗ページカウンタ、書き込み許可フラグ、失敗ページ表ポインタの領域を確保し、決められた仮想アドレス領域にマップする。カーネル36は、送信用および受信用の失敗ページカウンタ、書き込み許可フラグ、失敗ページ表ポインタの領域がページアウトされないように指定する。

【0075】生成されたプロセス11は、送信側、受信側ともに、データ転送に先だって、カーネル36のメモリ割り当て機能を利用して、失敗ページ表の確保と設定を行なう。このメモリ領域は割り込み処理172によりアクセスされるので、カーネル36は、失敗ページ表のメモリ領域がページアウトされないように指定する。例えば、Mac hオペレーティングシステムの場合では、vm_wir eシステムコールを使うことにより、失敗ページカウンタや失敗ページ表などのメモリ領域がページアウトされないように指定する。

【0076】次に、プロセス11は、このメモリ領域の先頭アドレスを送信用の失敗ページ表ポインタと受信用の失敗ページ表ポインタに設定する。プロセス11が受信プロセスであり、データの送信を行なわない場合は送信用の失敗ページ表ポインタを設定しなくてもよい。また、受信用の失敗ページ表ポインタは通信する送信プロセスの分だけ設定する。

【0077】送信側プロセス11（pid=n）は送信を開始する前に、送信側の失敗ページカウンタを「0」に初期設定し、送信側の書き込み許可フラグを「1」に設定する。さらに、受信プロセスの仮想空間にある自分のプロセス識別子に対応する失敗ページカウンタのアドレス（FT_BASE+2*（n+1）*4）に0x80000000を転送する。

【0078】そして、送信側プロセス11は、通信ハードウェア27の要求領域176にパラメータを書き込み、通信ハードウェア27にデータ12の転送を開始させる。

【0079】送信側ノードの送信部171での処理においてアドレス変換の失敗が起こると、送信側ノードの通信ハードウェア27の割り込み発生部172は、割り込みを発生する。送信側ノードの割り込み処理362の機能は、この割り込みを受け付けると、送信プロセス11の識別子を通信ハードウェア27の状態レジスタ（図示せず）からとりだす。割り込み処理362の機能は、送信側プロセス11の仮想空間にある送信用の失敗ページカウンタの書き込み許可フラグを参照する。

【0080】書き込み許可フラグの値が「1」であれば、割り込み処理362の機能は、まずその失敗ページカウンタの中のカウンタをインクリメントし、さらに、

10

20

30

40

50

失敗ページ表ポインタとカウンタの値から定まる場所
 ((先頭アドレス) + (カウンタ) * 4) にアドレス変換に失敗したページの仮想アドレスを記録する。書き込み許可フラグの値が「0」であれば、割り込み処理362の機能では、失敗ページ表に対する操作はなにも行わない。割り込み処理から復帰する際には、割り込み処理362の機能では、アドレス変換に失敗したページを飛ばすように指定して、送信側ノードの通信ハードウェア27の送信部171のデータ転送を再開させる。

【0081】一方、受信側ノードの受信部271での処理においてアドレス変換の失敗が起ると、受信側ノードの通信ハードウェア27の割り込み発生部172は、割り込みを発生する。受信側ノードの割り込み処理362の機能は、この割り込みを受け付けると、送信プロセス11の識別子と受信プロセス11の識別子とを受信側ノードの通信ハードウェア27の状態レジスタ(図示せず)から取り出す。割り込み処理362の機能は、受信プロセス11の仮想空間にある受信用の失敗ページカウンタのうち送信プロセスに対応するものの書き込み許可フラグを参照する。

【0082】書き込み許可フラグの値が「1」であれば、割り込み処理362の機能は、まず、その失敗ページカウンタの中のカウンタをインクリメントし、さらに、失敗ページ表ポインタとカウンタの値から定まる場所、例えば、失敗ページ表ポインタの値がRFTn_ADDR、カウンタの値がkであったとすると、RFTn_ADDR + 4 * kのアドレスにアドレス変換に失敗したページのアドレスを書き込む。

【0083】送信プロセス11は、何回かデータ転送を行なうと、データ転送の途中でアドレス変換の失敗が起きたかどうかを確認する。送信プロセス11は、送信側ノードの送信部171でのアドレス変換の失敗をチェックするために、自分の仮想空間にある送信用の失敗ページカウンタを読む。ここで、失敗ページカウンタ中のカウンタの値が「0」より大きければ失敗が起きたことになる。この場合は送信用の失敗ページ表ポインタの値から失敗ページ表の位置を特定し、そこからカウンタの分だけ失敗ページを読み出す。つまり、送信プロセス11は、失敗ページポインタの値がSFT_ADDR、カウンタの値がkであった場合には、SFT_ADDRのアドレスから4 * kバイト分読み出す。

【0084】送信プロセス11は、次に受信側での失敗をチェックするために、上述のリモートリード部571機能を使って受信プロセス11の受信用の失敗ページカウンタ及び失敗ページ表ポインタのアドレスを同時に読む。この領域はページアウトされないのでリモートリードの途中でアドレス変換の失敗は起こらない。ここで、失敗ページカウンタ中のカウンタの値が「0」より大きければ失敗が起きたことになる。この場合は、カーネル36が提供する遠隔ページアクセス761の機能を利用

して受信側ノードにある失敗ページ表をカウンタの分だけ読み出す。つまり、送信プロセス11は、失敗ページポインタの値がRFTn_ADDR、カウンタの値がkであった場合には、リモートリード機能によってRFTn_ADDRのアドレスから4 * kバイト分読み出すこととする。

【0085】なお、この実施の形態のマルチプロセッサシステムでは、送信プロセス11は、あらかじめ確保された失敗ページ表を越すようなデータ転送を行なってはいいけない。もしも、失敗ページ表を越すようなことが送信側ノードで起これば送信プロセス11が、受信側で起これば受信プロセス11がカーネル36の機能によって異常終了させられる。このような異常終了を避けるためには、ユーザプログラムで最初に確保する失敗ページ表のサイズを大きくすればよい。

【0086】以上説明したように、この実施の形態のマルチプロセッサシステムでは、複数のデータ転送に対する失敗確認と再送を一回で済ませることができる。これにより、別々のアドレスにある少量のデータを送るような場合に、失敗確認の回数を減らすことができるので、全体としてデータ転送に要する時間を第1の実施の形態の場合よりもさらに短縮することができる。

【0087】[第3の実施の形態] この実施の形態では、例えば、転送されるデータ12の最後の1ビットをデータの受信完了を示すフラグとして用いている。この実施の形態にかかるマルチプロセッサシステムの構成は、第1の実施の形態のものとはほぼ同一であるが、通信ハードウェア27の割り込み発生部172の機能として、次の2つの機能が追加されている。

【0088】すなわち、割り込み発生部172は、送信部171がデータ12を送信するときに、送信すべきデータのうちの最後のパケットとなったときに割り込みを発生し、受信部271がネットワーク10からデータ12を受信するときに、受信すべきデータのうちの最後のパケットとなったときに割り込みを発生する。

【0089】以下、この実施の形態のマルチプロセッサシステムにおける動作について説明する。カーネル36の割り込み処理362の機能は、第1の実施の形態で説明した割り込み処理に加えて、次のような割り込み処理を行う。なお、この割り込み処理以外の動作は、第1の実施の形態における動作と同一である。

【0090】図10は、この実施の形態において、送信側ノード3-1と受信側ノード3-nのカーネル36の割り込み処理362の機能が実行する割り込み処理を示すフローチャートである。割り込みが発生した場合、カーネル36の割り込み処理40の機能は、その割り込み要因について調べる(ステップS21)。

【0091】ステップS21で調べた割り込み要因が送信すべきデータのうちの最後のパケットとなったものである場合は、この例では、送信側ノード3-1で生じた

割り込みである。この場合、割り込み処理362の機能は、まず、送信側ノード3-1の送信失敗ページ表341を調べる(ステップS51)。そして、送信失敗ページ表341を調べた結果、送信部171の処理でいずれかのページにアドレス変換の失敗があったかどうかを判定する(ステップS52)。

【0092】ステップS52でいずれかのページにおいて、いずれかのページにアドレス変換の失敗があったと判定したときは、送信側ノード3-1の割り込み処理362の機能は、送信失敗ページ表341に最後のページについて(実際にはアドレス変換の失敗がなくても)アドレス変換の失敗があったものとして送信側ノード3-1の送信失敗ページ表341に記録する。そして、最後のパケットの送信を停止させて(ステップS53)、この割り込み処理を終了する。

【0093】一方、ステップS52でいずれのページにもアドレス変換の失敗がなかったと判定したときは、そのまま割り込み処理を終了する。この場合、最後のパケットは送信されることとなる。

【0094】ステップS21で調べた割り込み要因が受信すべきデータのうちの最後のパケットとなったものである場合は、この例では、受信側ノード3-nで生じた割り込みである。この場合、割り込み処理362の機能は、まず、受信側ノード3-1の受信失敗ページ表641を調べる(ステップS54)。そして、受信失敗ページ表641を調べた結果、受信部271の処理でいずれかのページにアドレス変換の失敗があったかどうかを判定する(ステップS55)。

【0095】ステップS55でいずれかのページにおいて、いずれかのページにアドレス変換の失敗があったと判定したときは、受信側ノード3-nの割り込み処理362の機能は、受信失敗ページ表641に最後のページについて(実際にはアドレス変換の失敗がなくても)アドレス変換の失敗があったものとして受信側ノード3-nの送信失敗ページ表641に記録する。そして、受信した最後のパケットを破棄させて(ステップS56)、この割り込み処理を終了する。

【0096】一方、ステップS55でいずれのページにもアドレス変換の失敗がなかったと判定したときは、そのまま割り込み処理を終了する。この場合、受信した最後のパケットは、受信バッファ22(メモリ13上にある)に書き込まれる。

【0097】以上のように割り込み処理を加えたことによって、アドレス変換に失敗したページを再送する場合には、最終ページも再送されることとなり、受信側ノード3-nでは、データの受信完了を示すフラグを含む最後のパケット(最終ページの最終パケット)を1番最後に受信することとなる。

【0098】以上説明したように、この実施の形態のマルチプロセッサシステムでは、受信側ノード3-nは最

後のパケットを必ず最後に受信することとなるので、例えば、データ中の最後の1ビットを受信完了を示すフラグとして用いるプログラムでも、正しく動作させることが可能となる。

【0099】[第4の実施の形態]図11は、この実施の形態にかかるマルチプロセッサシステムの機能構成を示す機能ブロック図である。このマルチプロセッサシステムは、第1の実施の形態のマルチプロセッサシステム(図1)とほぼ同様の機能構成を有するが、カーネル46が他のノードのカーネルからの要求を受けて、二次記憶装置18に存在するデータを送信先ノードに送信する遠隔データ送信1062の機能を有する点異なる。

【0100】以下、この実施の形態にかかるマルチプロセッサシステムにおける動作について説明する。この実施の形態のマルチプロセッサシステムの動作は、第1の実施の形態のものとはほぼ同一であるが、ステップS19の再送処理のみ異なる。また、ステップS19の再送処理においても、受信側ノード4-nでアドレス変換の失敗があった場合の処理は、第1の実施の形態のもの(図6(b))と同一である。

【0101】図12は、送信側ノード4-1でアドレス変換の失敗があった場合に、送信側プロセス11が実行する再送処理を示すフローチャートである。処理が開始すると、送信側プロセス11は、送信側ノード4-1の送信側失敗ページ表341を参照して、アドレス変換に失敗したページの仮想アドレスを計算する(ステップS31)。

【0102】次に、送信側プロセス11は、ステップS31で計算した仮想アドレスが示すページが遠隔ディスク(送信側ノード4-1以外にある二次記憶装置18)にあるかどうかを判定する(ステップS34)。

【0103】ステップS34で当該ページが遠隔ディスクにないと判定したときは、送信側プロセス11は、ステップS31で計算した仮想アドレスにアクセスしてページフォルトを起こさせ、カーネル46に当該アドレス変換に失敗したページを二次記憶装置18からメモリ13にページインさせる(ステップS32)。そして、送信側プロセス11は、そのページインされたページを再送する(ステップS33)。

【0104】ステップS34で当該ページが遠隔ディスクにあると判定したときは、送信側プロセス11は、カーネル46の遠隔データ送信要求1062の機能を利用して、当該ノードのカーネル31に対して遠隔データ送信要求を送る(ステップS35)。そして、送信側ノード4-1における処理を終了する。

【0105】次に、この遠隔データ送信要求を受信したノード(以下、遠隔ノードという)における処理を、図13のフローチャートを参照して説明する。遠隔ノードの遠隔データ送信1061の機能は、遠隔データ送信要求を受け取ると、その要求中に含まれている情報を元

に、遠隔ノードの二次記憶装置18から読み出すべきページを決定する(ステップS61)。

【0106】次に、遠隔データ送信1061の機能は、二次記憶装置18からステップS61で決定したページを読み出し(ステップS62)、受信側ノード4-nにそのページを送信する(ステップS63)。最後に、遠隔データ送信1061の機能は、送信が終了したことを要求元の送信側プロセス11に遠隔ノードのカーネル46が提供するプロセッサ間通信機能を使って通知する(ステップS64)。以上により、一連の再送処理を終了する。

【0107】以上説明したように、この実施の形態のマルチプロセッサシステムでは、遠隔ノードから直接データを送信するので、遠隔ノードから送信側のノードへのデータ転送を減らすことができ、データの再送の時間を短縮することができる。さらに、送信側ノードの負担を減らすこともできる。

【0108】【実施の形態の変形】上記の第1の実施の形態では、プロセス11が受信バッファ22のページインさせるためにカーネルが提供する遠隔ページアクセス機能761を利用したが、受信側ノードにページアクセス専用のスレッドを設け、そのスレッドを使ってページインさせてもよい。

【0109】このページアクセススレッドは要求メッセージを受けると、受信バッファ22の指定された仮想アドレスに直接アクセスすることによりページフォルトを起こし、カーネル36にページインさせる。この要求メッセージは受信プロセスの識別子とページインさせた仮想アドレスのリストを含む。なお、この要求メッセージは従来のカーネルが提供するのと同様のシステムコールを用いた通信機能を用いて送ることができる。

【0110】この方法を用いることにより、カーネルが遠隔ページアクセス機能761を持たない場合でも、第1の実施の形態と同様のマルチプロセッサシステムを構築することができる。

【0111】上記の第3の実施の形態では、最後のパケットを送信/受信する時に割り込みを起こし、その割り込み処理の中で最後のパケットを送信/受信しないようにする処理を行なっているが、最後のパケットを送信/受信する時の割り込みの代わりに、最後のパケットあるいは最後のページを送信/受信しないで捨てるような機能が通信ハードウェアにある場合も、同様のことが実現できる。

【0112】このような通信ハードウェアを使用した場合、最初にアドレス変換の失敗が起きた時の割り込み処理で、送信/受信失敗ページ表には最後のページも失敗したものとして記録し、さらに、通信ハードウェアに対して最後のパケット(あるいはページ)の送信/受信をしないで捨てるように要求する。これにより、最後のパケット(あるいはページ)は送信/受信されず、後で送

信プロセスが失敗確認した時には最後のページもアドレス変換に失敗したように見える。これにより、上記の第3の実施の形態と同様の効果を得ることができる。

【0113】上記の第1~第4の実施の形態では、ページング方式の仮想記憶をサポートするものに本発明を適用した場合について説明したが、セグメント方式やセグメントページング方式の仮想記憶をサポートする疎結合型のマルチプロセッサシステムにも本発明を適用することは可能である。

【0114】

【発明の効果】以上説明したように、本発明によれば、ノード間でのデータ転送を高速化することができる。また、不必要なアドレス変換や再送処理を行わなくてもよい。

【図面の簡単な説明】

【図1】本発明の第1の実施の形態にかかるマルチプロセッサシステムの機能構成を示す機能ブロック図である。

【図2】本発明の第1の実施の形態における失敗ページ表(送信失敗ページ表及び受信失敗ページ表)の構成を示す図である。

【図3】本発明の第1の実施の形態における失敗ページ表の仮想空間上の割付を示す図である。

【図4】本発明の第1の実施の形態にかかるマルチプロセッサシステムにおける動作を示すフローチャートである。

【図5】本発明の第1の実施の形態においてカーネルが実行する割り込み処理を示すフローチャートである。

【図6】本発明の第1の実施の形態においてプロセスが実行する再送処理を示すフローチャートであり、(a)は送信側で失敗があった場合の処理を、(b)は受信側で失敗があった場合の処理を示す。

【図7】本発明の第1の実施の形態にかかるマルチプロセッサシステムの動作例を説明する図である。

【図8】本発明の第2の実施の形態に適用されるページカウンタの構成を示す図である。

【図9】本発明の第2の実施の形態における失敗ページ表の仮想空間上の割付を示す図である。

【図10】本発明の第3の実施の形態においてカーネルが実行する割り込み処理を示すフローチャートである。

【図11】本発明の第4の実施の形態にかかるマルチプロセッサシステムの機能構成を示す機能ブロック図である。

【図12】本発明の第4の実施の形態においてプロセスが実行する再送処理を示すフローチャートであり、送信側で失敗があった場合の処理を示す。

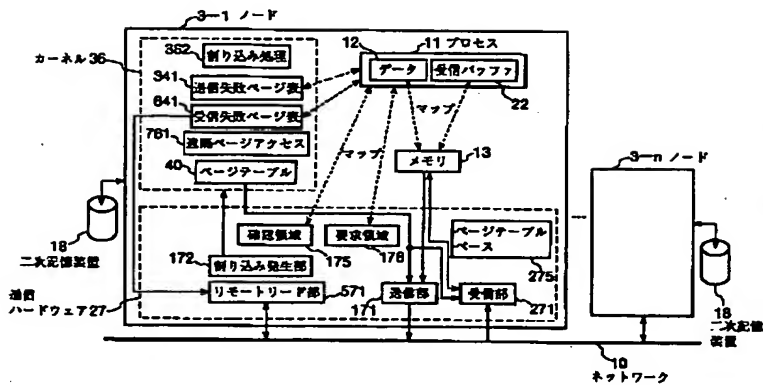
【図13】本発明の第4の実施の形態においてプロセスが実行する遠隔データ送信処理を示すフローチャートである。

【符号の説明】

25

3-1、3-n、4-1、4-n ノード
 10 ネットワーク
 11 プロセス
 12 データ
 13 メモリ
 18 二次記憶装置
 22 受信バッファ
 27 通信ハードウェア
 36、46 カーネル
 40 ページテーブル
 171 送信部
 172 割り込み発生部

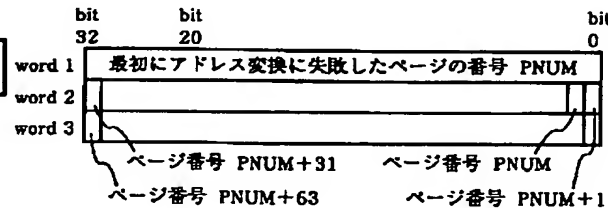
【図1】



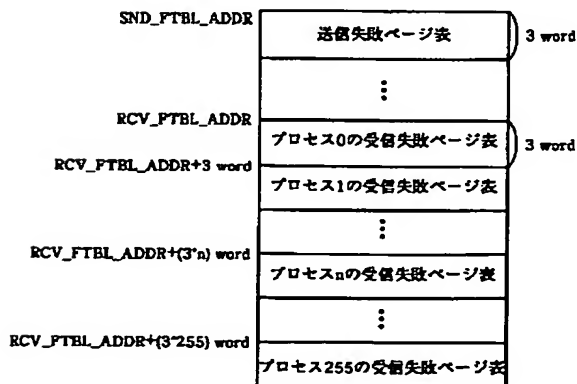
【図8】

書き込み許可フラグ(1ビット)

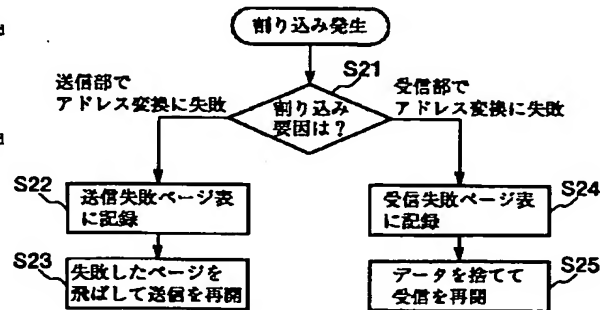
カウンタ(31ビット)



【図3】



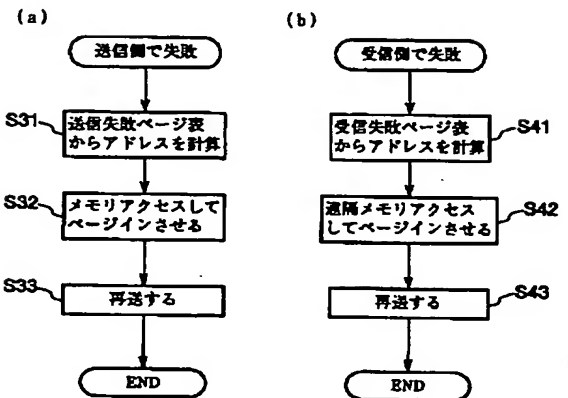
【図5】



26

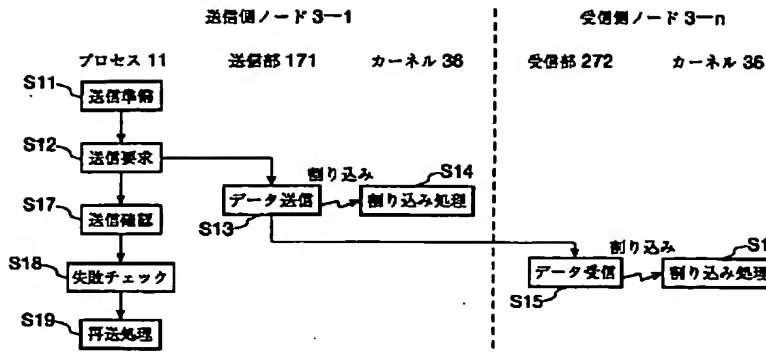
175 確認領域
 176 要求領域
 271 受信部
 275 ページテーブルベース
 341 送信失敗ページ表
 362 割り込み処理
 571 リモートリード部
 641 受信失敗ページ表
 761 遠隔ページアクセス
 10 1061 遠隔データ送信
 1062 遠隔データ送信要求

【図6】

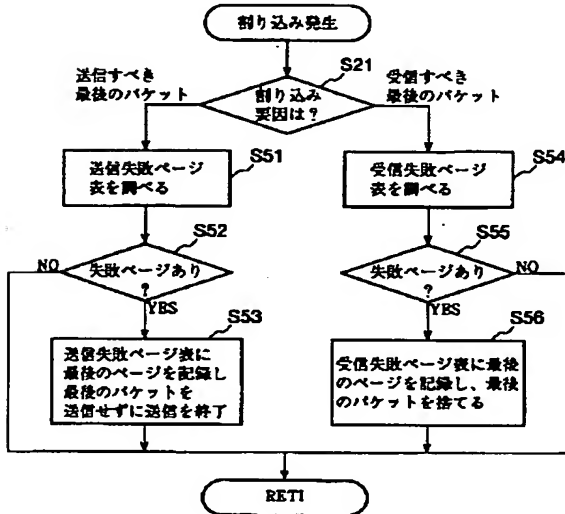


【図2】

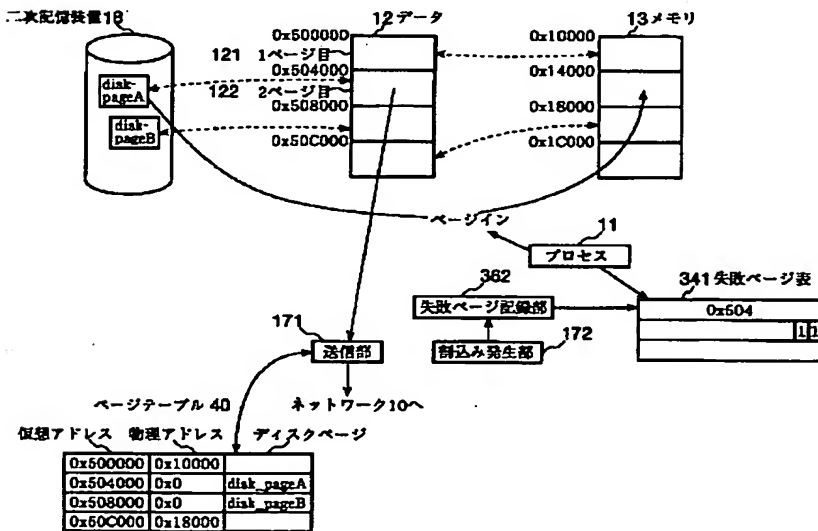
【図4】



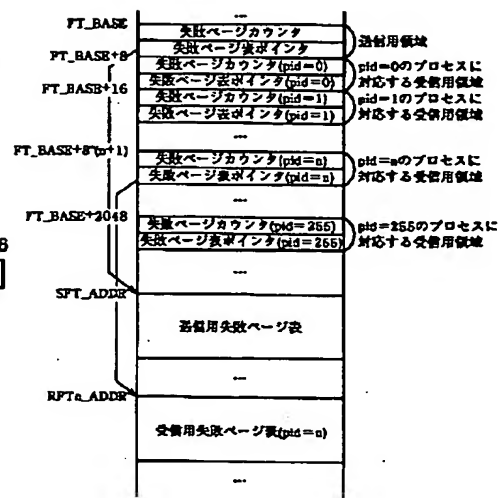
【図10】



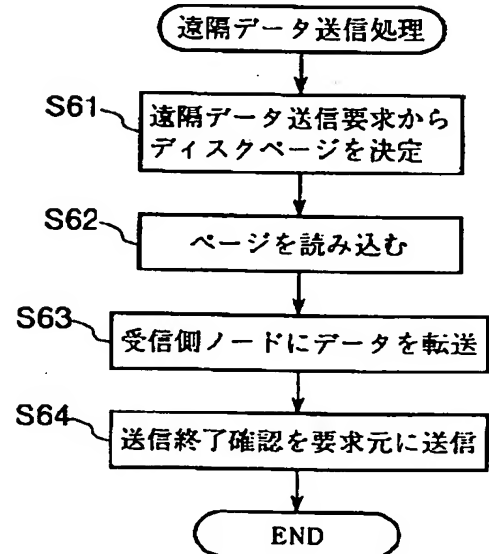
【図7】



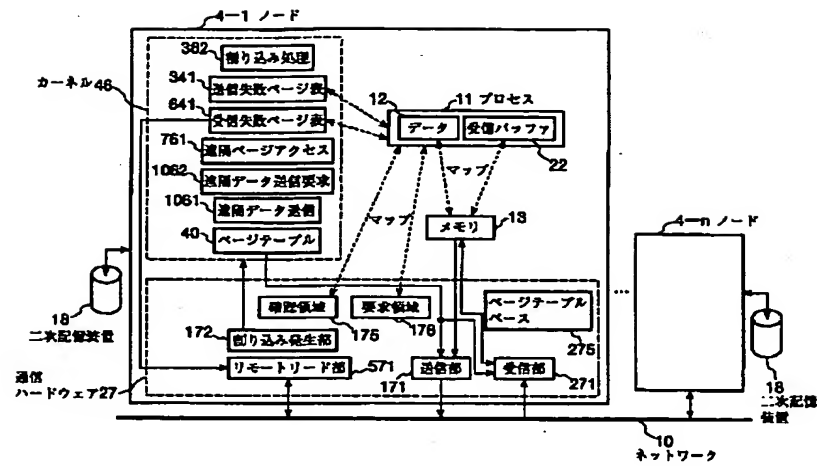
【図9】



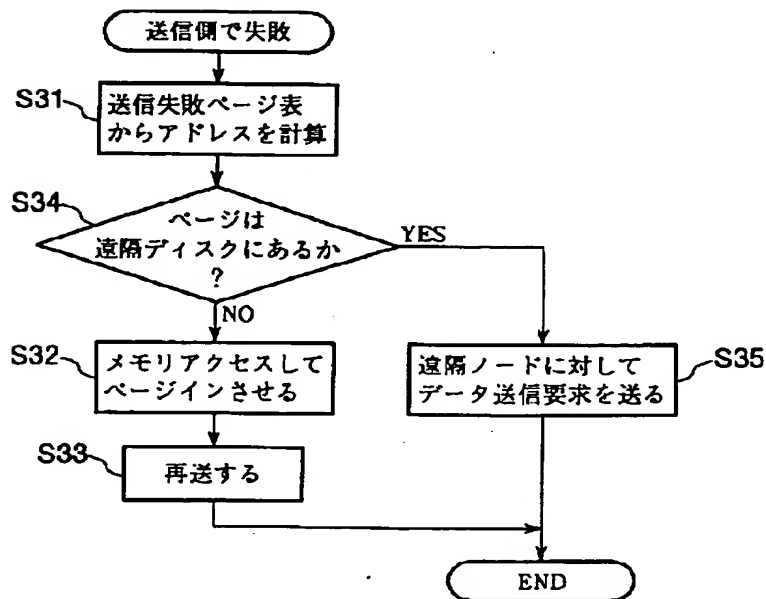
【図13】



【図11】



【図12】



フロントページの続き

- (56) 参考文献 特開 昭57-162164 (JP, A)
 特開 平7-64846 (JP, A)
 特開 平7-271739 (JP, A)
 特開 平9-212474 (JP, A)
 特開 平8-305667 (JP, A)
 特開 平7-262151 (JP, A)
 手塚宏史、外4名, 「ピンダウンキャ
 ッシュを用いたユーザレベルゼロコピー
 通信」, 情報処理学会研究報告V o 1.
 97 No. 76 (97-ARC-125),
 1997. 08. 22, p. 167-172
 石川裕, 「コモディティハードウェア
 を用いた並列処理技術」, 情報処理第39
 巻第8号, 1998. 08. 15, p. 784-791

- (58) 調査した分野(Int.Cl.⁷, DB名)
 G06F 15/16 - 15/177
 G06F 12/08 - 12/12
 G06F 13/00
 G06F 13/38 - 13/42
 J I C S T ファイル (J O I S)